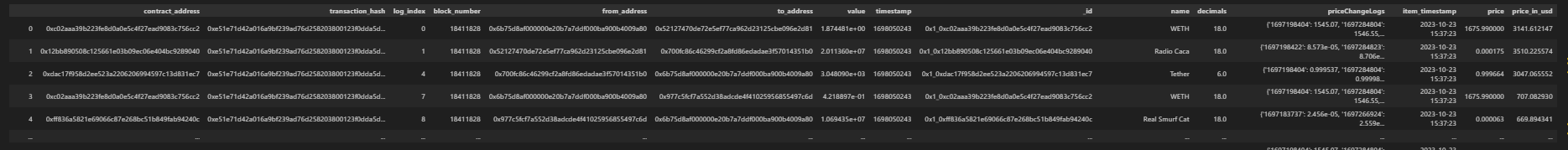
Datasets

**Dataset 1:** As of the problem 5’s scope, our group queried data from Centic database and re-synthesized them to create our dataset about the transaction history in term of token and wallet addresses. The retrieved data was saved in .csv format.

**Example:**



**Description:**

|  |  |  |
| --- | --- | --- |
| **Field** | **Data type** | **Description** |
| contract\_address | text | Address of interacted contract |
| transaction\_hash | text | The hash of the transaction |
| log\_index | bigint | The logs index position in the block |
| block\_number | bigint | The length of the blockchain in blocks |
| from\_address | text | Address of the sender |
| to\_address | text | Address of the receiver |
| value | double precision | The amount of token sent in this transfer. The value has been divided by token decimals. |
| timestamp | Int32 | The time when the block was mined that includes this transaction |
| item\_timestamp | String | Block timestamp in human readable format |
| name | String | Name of contract |
| decimals | Int32 | Token decimals |
| priceChangeLogs | Object | History of token price |
| price | Double | Token price |
| price\_in\_usd | Double | Token price in USD |

**Dataset 2:** Furthermore, to variate the data source, we continue to crawl information from <https://www.chainabuse.com/>. In this website, there are myriad of chain types such as: ETH (Etherium), BTC (Bitcoin), … where the scam reports can be collected. These reports can be in the kind of Phishing Scam, Sextortion Scam, Impersonation Scam, etc. Below is an example from json-dataset file.

{

"id": "605355bb-334f-423e-8be0-68d982394bc4",

"vote\_count": 1,

"category": "Rug Pull Scam",

"description": "AnubisDAO rug stealing over $60 million from investors. 0xB1302743ACF31F567e9020810523f5030942e211 holds over $40 million as of this scam report.",

"amount\_loss": null,

"reported\_addresses": {

"0xB1302743ACF31F567e9020810523f5030942e211": "eth",

"0x9fc53c75046900d1F58209F50F534852aE9f912a": "eth",

"0x872254d530Ae8983628cb1eAafC51F78D78c86D9": "eth",

"0x3069388B2C7e45a53EBB32c31bC3817304401D9E": "eth"

},

"reported\_domains": []

},

Despite of having chunk of report types, in general, they should contain below fields:

|  |  |  |
| --- | --- | --- |
| **Field** | **Data type** | **Description** |
| id | text | The index of report (it might be omitted in dataset but we still add it to make dataset be consistent and easy to retrieve). |
| vote\_count | int | The number of positive vote in this report |
| category | text | The type of report; e.g. Phishing Scam, Impersonation Scam, ... |
| description | text | The elaborated happened situation that the victim told |
| amount\_loss | float | The damage that the victim was suffered from. |
| reported\_addresses | text | Addresses of the causer, scammer.  It contained an sub-field in side to indicate which blockchain this address belong to.  For example: btc = Bitcoin |
| reported\_domains | text | Domain of the scammer |